DOCUMENT RESUME

ED 411 316 TM 027 382

AUTHOR Page, Ellis B.; Poggio, John P.; Keith, Timothy Z.
TITLE Computer Analysis of Student Essays: Finding Trait Differences in Student Profile.
PUB DATE 1997-03-00
NOTE 8p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; Elementary Secondary Education; *Essays; Evaluators; *Holistic Evaluation; *Scoring; Test Results; Test Use; Writing (Composition)
IDENTIFIERS *Residuals (Statistics)

ABSTRACT

Most human gradings of essays are holistic, or "overall." Therefore, Project Essay Grade (PEG), an attempt to develop computerized grading of essays, has concentrated most of its research on overall grading. It has successfully simulated human judges. However, since computer grading is less expensive than human grading, PEG has also explored the grading of traits within the essay (content, organization, style, mechanics, and creativity). PEG has found it possible to simulate multiple judges in grading such traits, but to make practical use of trait scores, it is important to discover how the traits vary within the students. In this study, 8 judges rated 495 essays on the 5 focal traits and the overall quality. Taking the holistic score as the overall essay value, researchers then studied the residuals of each trait from the holistic. These residuals turned out to be strikingly predictable. Using these traits and multiple judges, PEG programs may apparently supply diagnostic ratings together with the holistic scores. These may serve for the information of individual students and for use by teachers, school leaders, and test researchers. (Contains 4 tables and 18 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 411 316

Computer Analysis of Student Essays: Finding Trait Differences in Student Profile

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

John Poggio

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Ellis B. Page
John P. Poggio
Timothy Z. Keith

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM027382

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in Chicago, IL
March 1997

BEST COPY AVAILABLE

HANDOUT for AERA NCME Symposium on Grading Essays by Computer.
1997 Meetings of AERA Div.D and of the NCME. Chicago, March 27, 1997

Computer Analysis of Student Essays: Finding Trait Differences in the Student Profile

Ellis B. Page — *Duke University*
John P. Poggio — Kansas University
Timothy Z. Keith — *Alfred University*

Most human gradings of essays are Holistic, or "overall". Thus, Project Essay Grade (PEG) has concentrated most of its computer research on such overall grading, and has had success in simulating multiple human judges. However, since computer grading is much less expensive than human grading, PEG has explored the grading of *traits* within the essay (here content, organization, style, mechanics, and creativity). Two years ago, PEG found it possible to simulate multiple judges in grading such traits. However, to make practical use of Trait scores, it is important to discover how such Traits vary within the student. In this work, 8 judges rated 495 essays on those five traits and the overall quality. Taking the Holistic as the overall essay value, we then studied the *residuals of each trait from the Holistic*. Such residuals turned out to be strikingly predictable. Using such traits and multiple judges, PEG programs may apparently supply *diagnostic* ratings, together with the Holistic scores. These may serve for the information of individual students and for uses by teachers, school leaders, and test researchers.

Recent studies have demonstrated that computers can grade essays better than 2 or more judges (where quality is determined by predicting ratings by larger groups of judges (*cf.* Page, 1994; Page & Petersen, 1995). Most studies of esssay grading have been limited to overall ("holistic") ratings, and for good reason: Human ratings are already expensive, and any diagnostic description, such as traits within the essay, would be prohibited by huge extra costs.

Computer ratings, however, are costly only for the norming sample. And this sample could be a tiny percentage of the student participants, with per-student costs very low for the larger population.[1] Thus, we may now explore offering some diagnostics of the essay, together with the overall rating. Here we describe recent attempts to simulate 8 human judges not only in their Holistic ratings, but in their ratings of five traits considered important in essays (Page, Keith, & Lavoie, 1996). And we especially describe brand new experiments to make useful diagnoses *within* the student essay.

Recent PEG grading of essay traits

For some of our recent work since 1992, we have used essays which were collected for a federal research: the Writing Assessment of the National Assessment of Educational Progress (NAEP). We especially have used the 12th-grade essays for the studies of 1988 and 1990. Each NAEP essay already had received one holistic rating, and was partly computer-ready. We added more ratings from qualified judges, to meet our own many needs of the PEG research.

It is interesting to reflect on the extra benefits we might gain from using the more efficient and economical computer grading. One of these would surely be to provide more feedback to the students, teachers, and researchers. So why not consider simulating the judges' ratings of various traits *within* the essay? Here is one acceptable list of such traits:

Content
Organization
Style
Mechanics
Creativity.

We used these, and we included Holistic again, partly to study its interactions with the traits, and to put all of these in proper perspective.

Eight qualified judges rated these characteristics for the 495 essays in the NAEP essays of 1988. This means each judge logged 2,970 scores. But how should the traits be presented on their rating sheets? After all, we have little evidence about how judges react to multiple traits. Should we present the traits first, so that the judge will consider these first? Or will that distort the Holistic outcome?

These and other questions were solved by a complex Latin square design (described in Page et al., 1996).

We also decided *not* to "train" the judges. Test companies have good reasons to train them, but these reasons did not hold for the present research -- where we were more interested in *sampling* the English teachers than in directing them.

As noted, each judge contributed many ratings, Holistic and 5 traits for each of 495 essays. Some of the judge results are analyzed in Table 1.

[Table 1]

In Table 1 we observe that Holistic and the 5 traits differed in their judge agreement. Perhaps surprisingly, the highest agreement was on the Holistic rating.

Also Table 1 shows the average correlations for the pairs of judges (Col. 2), then for 3-groups, 4-groups, and 8-groups of

judges. As the judges rise in number, the agreements between their groups rise as well. With the 8-group, we reach reliabilities ranging from .87 to .93. Table 1 provides us with useful targets for our predictions (since it is rare that predictions exceed the reliability of the criterion).

[Table 2]

In Table 2, we see in the first column the average Mult-R's generated for Holistic and for each trait, within the Formative samples of about 400 essays each, for each of 100 random trials. We see that these averages range from .92 (for Content) to .86 (for Mechanics). These Mult-R's correlate .72 with the typical agreement between human raters. And the Cross-validations correlate .67 with these judge agreements. We remember that the higher judge agreements make for a more reliable criterion for regression, thus helping to increase the apparent power of the regression.

We also see that there is a high relation between the Mult-Rs and the Cross-validations (.97 in this tiny sample). Interesting also is the relative shrinkage in the Cross-validations. The largest shrinkage was with Style (.064), and the least shrinkage with Content (.27).

Of course, such differences in power (in Mult-R and in Cross-validation) may come from many causes other than the trait itself, or the reliability of the judge averages. They may also reflect relative strengths in the PEG variables used to make these simulations. Then too, there is the possibility of random fluctuations in human judgments, especially those given to extreme essays with unusual properties.

[Table 3]

Now we come to the content of Table 3, summarizing the direct comparisons of the computer program (PEG-7) and the various groupings of the human judges. To generate this table, we began with the first column of Table 1: the average correlations between human judges. From those, we used the Spearman-Brown prophecy formula to generate the first 4 columns of Table 3.

These three columns show how well we would expect judges, one or more, to predict the ratings of 8 other judges. Thus, using the basic average correlation between the individual human judges (.61, from Table 1), we forecast how well one judge would predict 8 judges (.75), and this number appears for Holistic in Table 3. The rest of the predictions are seen. We observe that a typical group of 4 judges would predict 8 judges at .89. But remember the practical world of essay rating: Our major concern must be with *two* human judges, since two are all that can be afforded (except for rare cases) in large scoring programs. Thus, in Table 3, the "*2 jud*" column is high-lighted, to keep that contrast in mind. But predictions for 3 judges and 4 judges are also tabled.

The central comparison is with the 5th column, PRED: the record of performance of the 600 Cross-validations generated by PEG, and the average agreements of PRED with the actual judge ratings on the 600 Test samples. In all cases, PRED is ahead of the 2-judge level, in most cases strikingly so.

To clarify these results, we have made comparisons not only with the first four columns, but also with prophecies for 5 and 6 judges. The overall PRED results are expressed in the last column, "PEG performance as N of judges." In this final column, we see the worst performance was that for Holistic and Style, yet even these surpassed 2 judges, and were slightly ahead of three.

Still more surprising are the PRED correlations with Content and Creativity. Creativity reached the 6-judge accuracy, and Content clearly passed its 6-judge comparison.

In short, the PEG approach has apparently moved strongly into grading traits within a set of essays.

Moving from Research to Feedback

INTERESTING AS THEY ARE, these findings do not yet provide us with all the applications we might wish.

How about practical information for the diagnostic reporting of the student performance? Since there is a high correlation between the Holistic and *Trait* scores, how can we find out where the student's own trait scores may be compared with one's Holistic performance?

In our latest work with Traits, we have addressed this within-student variation.

Reasons for early pessimism

We know the usual problems of within-subject ratings: What some call the *halo effect* needs to be factored in. That is, if Johnny is at the bottom of a class in Holistic, he may well be at the bottom in *Content, Organization, Style, Mechanics*, and *Creativity*. What use is there in telling Johnny, or his Teacher, such information?

Furthermore, with just one or two graders, the subtler within-student differences are not going to be well-measured, even if they are present.

Is it possible that, despite the high correlations of these traits with Holistic and with each other, we may achieve some solid and useful discrimination within student? This became a most interesting question, particularly given the rare data with 8 judges for each essay, and given the astonishing opportunities presented by computer grading.

Judges are at the center

With these NAEP data, we have both the Traits and a Holistic score. (In some essay datasets, the Holistic is not given by the judges directly, but is inferred by factor analysis.) Here we placed the Judges at the center of our study: We took the "Holistic" at face value, as representing a Judge's own opinion about overall value in an essay. In brief, we let the *Judge* decide what is "important" in the essays.

Then the question is whether there is any reliability in the residual variance, once the Holistic is subtracted from a Trait's score.

Would there remain any important discrimination in such traits?

Regression analysis of the Trait residuals

First, we developed an Average score for each essay, on

each trait. This was done by standardizing the 8 judge ratings for each trait, and then subtracting from each trait the standardized judgment for Holistic (also across 8 judges).

Then we performed a Linear Regression analysis on these Trait residuals. The principal results are seen in Table 4.

[Table 4]

The first thing we notice in Table 4 is the high levels of the Multiple Regressions, from .30 for Organization, up to .69 for Mechanics. Needless to say, with a short list of composite predictors, all of these are at an extremely high significance level.

We see also that the Standard Deviations of the *predicted deviations* are roughly correlated with the Mult-R's. The next two columns are the respective Minima and Maxima of these predicted scores. Not surprisingly, Mechanics furnish the largest deviations from Holistic.

Should we therefore conclude that most student papers are indeed most deviant in Mechanics? Not necessarily. After all, this dominant deviation of Holistic may well be caused by Judge behavior, rather that student behavior. It might be that the Judges grouped the others as being more similar, and closer to *Holistic*, and viewed Mechanics as more independent. They may also have been more censorious about errors in mechanics.

So, how should these results be treated? What should be the "feedback" or advice to educators and students? This is a question which deserves a practical answer, or a policy answer, as much as it does a statistical answer.

We are working on these questions. In any case, in Table 4, the power of the discrimination among these residuals was startling to us. Here it seems evident, despite the reservations and doubts we felt before this analysis, that *all* of the Trait scores yielded contrasts which were remarkably significant.

Strengths and weaknesses of this experiment

By their nature, most experiments are limited in generality. All experiments must work within given samples, and must use the tools at hand. Let us look again at some aspects of this experiment:

The essay sample. These 1988 NAEP essays were collected to be a stratified random sample of senior students in American high schools. They were written by students who responded to a particular question (the "Recreation Decision"). Students had no particular incentive for doing well. They wrote by hand, and their essays were later entered by typists under special instructions.

But the national sampling of High School & Beyond was about as good as we've had in the U.S. We also have other evidence that correlations are strong between ratings for hand-written and for machine-entered essays, so that should not matter too much. (All of these 8 trait judges worked from clear printed copy.) And for essay type, there is broad generality in the "persuasive" genre, and its importance in education of citizens.

More broadly, we now have a background of very different samples of writing judged by PEG, all with considerable success. Some of these have been much younger groups (junior high level), and others have been older and more advanced (the advanced college students taking their ETS Praxis essays; and most recently the still more advanced students taking the GREs.

The human raters. Some of the 8 raters were English teachers with broad experience. All had bachelors, and most had advanced degrees. All were in the top 5%, or higher, of the national intellectual pool. These compare well with those usually employed in large testing programs to rate papers.

The trait ratings. As noted in Table 1, the judges were not as much in agreement on the traits as they were on the overall (Holistic) ratings. In a large essay grading program, judges are typically "trained", but that was not done here because we wanted the broader generality of opinion about what constitutes these traits.

How does the judge agreement (or lack of it) affect the comparative performance of PEG? Two ways: First, a slightly higher agreement makes for a more "reliable" group opinion, which may increase the Mult-R and the Cross-validation. On the other hand, the way we define PEG "performance" here is in the number of judges to reach that PEG level. Judge lack of agreement may make it easier for the computer to pass this level of 1 judge, 2 judges, and so on, in predicting the larger group. (Surely, if any judges are absurdly high in agreement, it would be technically impossible to surpass them.)

In the future, we will probably experiment some with "trained" judges for traits (as we did for the Holistic ETS ratings for both the Praxis and GRE essays), and may have better knowledge about this.

The PEG program. In the last three years, we have made many changes in the working program, and the accuracy seems to have been improving. In this experiment, we can say with confidence that PEG rated all traits much better than the 2-judge level. And we see evidence that all traits share some predictors with other traits (though the relative weighting of these predictors will often be different for different traits).

Still, there may be some differences between traits because PEG itself may handle one trait better than others, perhaps from having special variables for some traits, but lacking others. In that case, we might in the future find the order of success altered across the traits.

The statistical program. Our new PEG statistical methods provide us with number-crunching programs which greatly speed our research. Earlier, we have depended on just a handful of randomly sampled replications, in our effort to measure the true effects and to refine our productivity. With our new programs, we can easily focus on the true shrinkage, in ways much faster and easier than those of standard statistical programs.

Will computer essay grading be accepted?

It is one thing to show the apparent feasibility of such methods. It is another to change habits of thinking about essay tests. And there are inevitable objections to something this new, which may seem to threaten the more ancient approaches to essay grading.

Philosophical and technical objections may be grouped into three kinds: *humanist*, *defensive*, and *construct*.

1) *The humanist objections*: Humanist critics believe that only a reasoning human being can make judgments about essay grades. And since the computer is not a human being, it is ridiculous to consider the computer grading essays. The idea should be dismissed.

This argument was much more common at the dawn of the computer revolution. Alan Turing gave a famous response with his "difference game": There were two doors, with a human behind one, a computer behind the other. If you could not tell whether human or computer was answering, then the computer won.

But -- as we have seen in our tables -- such a "difference game" would now give all victories to the computer.

2) *The defensive objections*: What if we have a mischievous or hostile student? Can't such students embarrass the program by submitting foolish answers in the "correct" form?

All of the essays so far graded by PEG have been "good-faith" essays. We have yet to do research on "bad-faith" essays, because we have had none to work with.

For the immediate future, in *large* essay programs, we would hope to run the PEG program in parallel with one human judge, who can easily check for improper, off-beat, or off-topic essays.

In this way, all will be reassured that human raters are present. And the computer cost may still be less than that of a second judge. The resulting quality of data (perhaps including such trait ratings as here described) would be far superior.

In the slightly longer run, we would hope to make checks which would guard against most such bizarre essays, and set them apart for human examination. (Such third-party evaluation is not new. It is now done with perhaps 5% of essays in current large programs.) And we would welcome the research challenge.

3) *The construct objections:* Some would say, despite the evidence of superior ratings, that such programs are looking at the "wrong things". Such critics would dismiss the use of "proxes" and would insist on "trins". And in their concept of trins, only human judges could suffice.

Yet let us think about the human judges in service, now doing these ratings. Does any judge really know the "trins" of any other? Their agreements are rather low, so they evidently are not working with just the same trins.

Also consider: In every large rating program, some judges are not invited to continue in future sessions. Why not? Because they did not agree enough with the other judges. This is virtually the sole evidence we have of the quality of their judgment. Such a judge may be the one, "true" judge of quality -- but we will not know, because we have no way of knowing. Still, that judge will no longer be used. We insist on substantial correlations between our judges.

How can such a test satisfy us about the human rater, but not about a computer system? Why can't we apply the same standard to the computer? Then it would win with ease. It would *always* be invited back -- and given preferred status.

In conclusion

That ancient test, the essay exam, is apparently increasing in importance, even within large objective testing programs, and such essay tests are now mandated by many large state and city school systems.

Yet even with two raters (the most common number), such tests have poor reliability for individual student decisions, and are virtually useless for other psychometric or research use.

In recent work, Project Essay Grade has evidence of matching the Holistic performance of multiple human judges. Potentially, it may provide useful data for comparisons across groups, schools, and years. A blind test, conducted with ETS, has shown that the computer can assign ratings to new essays not seen before, and can correctly forecast the group judgments better than even 3 human judges.

In this latest work, we have analyzed 495 essays and have simulated the 8 human judges better than would 3 other human judges. For the first time, we have also used this powerful system not simply for Holistic ratings, but for five traits commonly accepted as fundamental to essay quality: Content, Organization, Style, Mechanics, and Creativity. Powerful statistical programs have allowed us to run 100 new formative and test programs to zero in on the accuracy of our predictions, across the long perspective.

New research is underway for other aspects of such essay grading: for increasing our accuracy still further, for studying outliers in our predictions, and for accommodating the PEG system to the needs of adaptive testing. Also under study is the possibility of providing helpful assistance for the classroom teachers of America. There seems to be a large field opening up for applications, and for expansion of theory.

References and Working Papers

A jay, H.B., Tillett, P.I., & Page, E.B. (1973). *Analysis of essays by computer* (AEC-II). Final Report to the National Center for Educational Research and Development. Washington, D.C.: Department of Health, Education, and Welfare.

Daigon, A. (1966). Computer grading of English composition. *English Journal, 55*, 46-52.

Keith, T.Z. (1994). Relative validity of computer and human ratings: LISREL analysis of the PEG data. Paper presented at Annual Meeting of the North Carolina Association for Research in Education (March 18, Greensboro, NC).

Page, E.B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238-243.

Page, E.B. (1967a). Grading essays by computer: Progress report. Proceedings of the 1966 Invitational Conference on Testing. Princeton, N.J.: Educational Testing Service.

Page, E.B. (1967b). Statistical and linguistic strategies in the computer grading of essays. *Proceedings of the Second International Conference on Computational Linguistics, 34.* (Abstract.)

Page, E.B. (1993, January). New computer grading of student prose, using a powerful grammar checker. Paper presented at the Annual Meeting of the North Carolina Association for Research in Education. Greensboro, N.C.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*(2), 27-142.

Page, E.B., Fisher, G.A., & Fisher, M.A. (1968). Project Essay Grade: A Fortran program for statistical analysis of prose. *British Journal of Mathematical and Statistical Psychology, 21*, 139. (Abstract).

Page, E.B., Truman, D.L., & Lavoie, M.J. (1994). "Teacher's Helper": Proposed use of Project Essay Grade for the English classroom. Annual Meeting of the South Atlantic Modern Language Association, Baltimore, MD, Symposium, November 11.

Page, E.B., & Paulus, D.H. (1968). *The analysis of essays by computer* (Final Report to the Bureau of Research at the U.S. Office of Education). Washington, DC: U.S. Department of Health, Education, and Welfare.

Page, E.B. (1985). Computer grading of student essays. In Husen, T., & Postlethwaite, N. (Eds). *International Encyclopedia of Educational Research*. Oxford, England: Pergamon. Pp. 944-946.

Page, E.B., & Petersen, N.S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, March, 76(7), 561-565.

Page, E.B., Tillett, P.I., & Ajay, H.B. (1989). Computer measurement of subject-matter essay tests: Past research and future promise. *Proceedings of the First Annual Meeting of the American Psychological Society*, 1, 39. (Abstract)

Page, E.B., Truman, D.L. & Lavoie, M.J. (1994). "Teacher's Helper": Proposed use of Project Essay Grade for the English classroom. *Annual Meeting of the South Atlantic Modern Language Association*, Nov. 11, Baltimore, MD.

Paulus, E.H., McManus, J.F., & Page, E.B. (1969). Some applications of natural language computing to computer-assisted instruction. *Contemporary Education, 40*, 280-285.

Rumelhart, D., McClelland, J., and the PDP Research Group (1986). *Parallel Distributed Processing*. Cambridge, MA: MIT Press.

NOTES

1. While computer ratings are much less expensive than human ratings, there may be a period in early large essay programs when the computer will run in parallel with *one* human judge, as backup against "bad faith" essays (see later).
2. *Cf.* Page & Petersen, 1994. ETS participation in this experiment does not mean that ETS will introduce such methods into its large essay programs.
3. During this same period of exploring large programs, PEG also investigated the simulation of teacher grades, with classroom samples from North Carolina and Connecticut (Page, Truman, & Lavoie, 1994).
4. Such "checking" procedures would help in surveying for off-beat or inappropriate essays. These might be rated, but then sidelined for human overview.
5. This algorithm would be very time-consuming with any standard statistical package.

For correspondence or reprints, please note the following :

Dr. Ellis B. Page.
213 West Duke Bldg., Duke Univ.,
Durham, NC 27708
or call Duke Voice Mail:
(919) 660-3084
or write me:
Project Essay Grade
110 Oakstone Dr.
Chapel Hill, NC 27514-9585
or E-mail to:
EBPAGE@ACPUB.DUKE.EDU

TABLE 1
Intercorrelation and reliability of human raters

Judged Variable	Correlations between single or groups of judges: 1 jud	2 jud	3 jud	4 jud	8 jud	
Holistic	0.61	0.76	0.82	0.86	0.926	Hol
Content	0.52	0.68	0.76	0.81	0.897	*Con*
Organization	0.45	0.62	0.71	0.77	0.867	*Org*
Style	0.49	0.66	0.74	0.79	0.885	*Styl*
Mechanics	0.46	0.63	0.72	0.77	0.872	*Mec*
Creativity	0.53	0.69	0.77	0.82	0.900	*Cre*

Note: The first column is the average correlation between single judges for the trait listed to the left. The second column is the typical corr. between two pairs of judges. The "3 jud" column is the corr. between groups of 3 judges, and this is continued for four judges. Finally, we show the corr. between 2 groups of 8 judges; and this is also the reliability of the total 8-judge group.

BEST COPY AVAILABLE

TABLE 2
PEG Multiple -R's, Cross-validations, and predictions of single judges

Judged Variable	PEG Mult-R	PEG CrossVal	PEG corr Av. 1-jud	Avr .corr. bet 1-jud	
Holistic	0.908	0.876	0.712	0.61	Holl
Content	0.917	0.890	0.678	0.52	*Cont*
Organization	0.878	0.841	0.602	0.45	*Organ*
Style	0.881	0.817	0.607	0.49	*Style*
Mechanics	0.856	0.796	0.576	0.46	*Mech*
Creativity	0.913	0.881	0.673	0.53	*Creat*

Note: The first column shows the average Mult-R reached across 100 random replications, each using 50 selected variables. The second column shows the cross-validations, applied across about 100 random test essays. The third column is the average correlation between the PEG prediction and the individual human rater. This is much larger than the typical interjudge agreement, shown in the 4th column.

TABLE 3
PEG: Prediction of 8 Judges by 1, 2, 3, 4 Judges, and by the computer's PRED

Judged Variable	*Prediction of 8 judges by:* 1 jud.	2 jud.	3 jud.	4 jud.	PRED	*PEG performance as N of judges*	
Holistic	0.75	0.84	0.87	0.89	0.876	*3*	Holistic
Content	0.68	0.78	0.83	0.85	0.890	*6+*	*Content*
Organization	0.62	0.73	0.79	0.82	0.841	*4+*	Organiz.
Style	0.66	0.76	0.81	0.84	0.817	*3+*	*Style*
Mechanics	0.63	0.74	0.79	0.82	0.796	*3*	*Mechanics*
Creativity	0.69	0.79	0.83	0.86	0.881	*6-*	*Creativity*

Note: These predictions of 8 judges by lesser numbers were generated from the Spearman-Brown prophecy formula. The prediction of 8 judges by PEG, however, is shown in the column for PRED, and comes directly from the cross-validations shown in Table 2. The final column combines these earlier columns. This final column (on "performance") shows the power of PEG in comparison with 3 judges or more -- even with 4, 6, or more judges.

TABLE 4
Predicting the Deviation of Trait Scores Around the Holistic Scores

Deviating Trait	*Data from the Predicted Residuals for Five Rated Traits* Multiple-R	St. Dev.	Min.	Max.	Trait
Content	.48	.32	-.90	1.15	*Cont.*
Organization	.30	.34	-1.12	1.00	*Organ.*
Style	.46	.32	-.95	1.25	Style
Mechanics	.69	.60	-1.35	2.40	*Mech.*
Creativity	.45	.38	-.98	1.27	Creat.

\wav\trait04.97

Note: These data represent the first effort to test the *predictability* of the Trait Residuals around the Holistic scores. It is notable that Mechanics was most predictably deviant from Holistic, though all were significantly separate from Holistic. Important questions remain about how the results should be communicated to writer and/or school personnel. These represent 495 essays, as rated on Holistic and the 5 Trait scores by 8 qualified, independent judges.

BEST COPY AVAILABLE